Enter the Dragon's Dojo: Computing Interview Questions for Human Experts to Classify Database Keys as Meaningful

Zhuoxing Zhang
The University of Auckland
Auckland, New Zealand
zzha969@aucklanduni.ac.nz

Sebastian Link
The University of Auckland
Auckland, New Zealand
s.link@auckland.ac.nz

ABSTRACT

Database keys enable the effective and efficient processing of core data management tasks and downstream applications. We introduce a tool, called Dragon's Dojo (DD), that provides computational support for domain and database experts to systematically analyze which minimal keys should be specified on a given schema. DD generates yes/no questions for experts to answer: A "no" means the corresponding key is meaningful, while a "yes" means the key is meaningless. The underlying methodology forms the human-centered counterpart that augments and complements the discovery of database keys from data. We will demonstrate how DD i) generates questions based on different strategies to traverse the space of possible questions, ii) interactively minimizes the number of questions required, and iii) enables domain and database experts to work together using translations between Armstrong samples and sets of minimal keys. The audience competes in a game where they need to find the set of minimal keys in a fixed scenario.

CCS CONCEPTS

• **Human-centered computing** → *Interaction design*; • **Information systems** → **Relational database model**; **Integrity checking**; **Inconsistent data**; **Database utilities and tools**.

KEYWORDS

Algorithm, Database key, Experiment, Requirements acquisition

ACM Reference Format:
Zhuoxing Zhang and Sebastian Link. 2026. Enter the Dragon's Dojo: Computing Interview Questions for Human Experts to Classify Database Keys as Meaningful. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article x, 4 pages. https://doi.org/x

1 INTRODUCTION

Database keys are combinations of columns whose values identify every record in a table uniquely [3]. They ensure that every real-world entity is represented uniquely within the database, and are therefore critical for conceptual, logical and physical database design, schema management, data cleaning and integration. We use the symbol R to denote the set of keys that ought to be satisfied by every relation that expresses a real-world database instance of the application domain. We refer to keys in R as *rules* or *meaningful keys*. In practice, R is typically different from the set S of keys that are currently perceived as rules. On the one hand, all keys in $R - S$ are false negatives: actual rules not perceived as such. On the other hand, all keys in $S - R$ are false positives: keys that are incorrectly perceived as rules. There are many reasons for false positives and negatives: new datasets may have not been analyzed yet, existing datasets may have been only analyzed partially, application requirements evolve, resources are not sufficiently available. False positives mean that some real-world data is incorrectly prohibited from entering database instances, while false negatives mean that data duplication is incorrectly permitted in database instances. Unless $S = R$, update maintenance and query processing do not perform in terms of the results they return and the speed they operate.

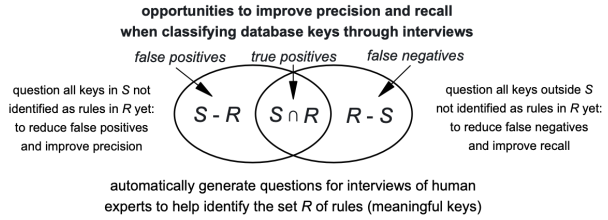


Figure 1: Opportunities to improve precision and recall of classifying database keys as meaningful and meaningless

Data profiling has been proposed to discover classes of database constraints from given data, by algorithmic means [1]. These classes include keys. However, any output of such algorithms cannot distinguish between those that are meaningful and those that only hold accidentally on the given instance. Hence, the results of key discovery may be viewed as the set S above. Even though there have been attempts to return more meaningful and fewer meaningless keys [2, 4, 5, 8, 9], results of algorithms are only recommendations, and final decisions can only be made by human experts.

We propose the first tool, called Dragon's Dojo, that directly supports domain and database experts in their task of identifying the set R of meaningful database keys, only given a set of columns and not utilizing any data. The approach is focused on human interaction, automatically generating Yes/No interview questions for them to answer. The process can use and classify data profiling results into meaningful and meaningless, which is only possible for human experts. As illustrated in Fig. 1, the tool automatically generates questions that exhaust all potential keys, thereby guiding human experts in eliminating false positives and negatives.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/x

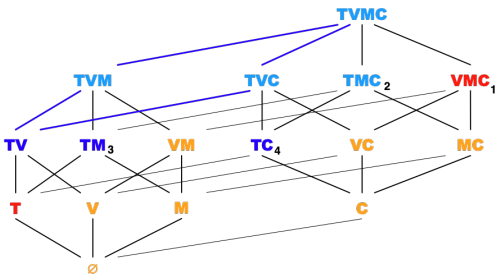


Figure 2: Traversal Order for Running Example

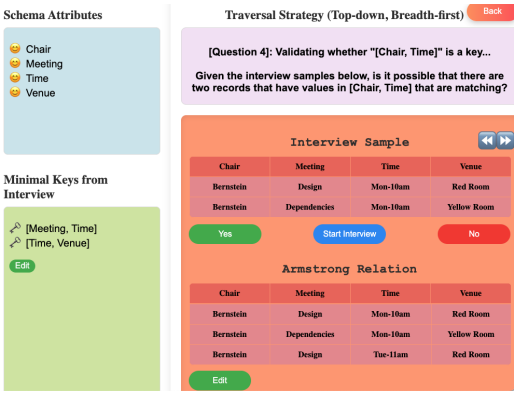
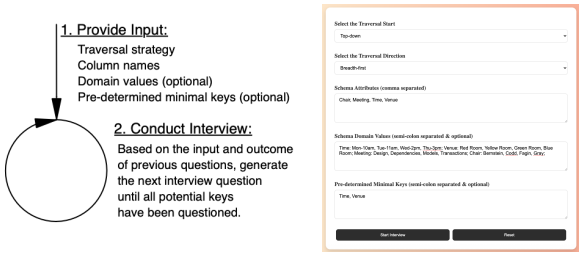


Figure 3: Interface of Dragon's Dojo with Running Example

Example 1. *We use a simple* Calendar *application that schedules a* M*(eeting) at some* T*(ime) and* V*(enue) with a person acting as* C*(hair). Based on a strategy the expert team chooses for traversal, the tool generates for each potential keys a Yes/No question the team collectively answers. Each answer determines how the search space is pruned and traversed. For the answers given, the set R of minimal keys is returned with the fewest questions required for the traversal strategy chosen. Assume profiling returns* $S = \{VT, CMT\}$*, the key* VT *is known as meaningful but the team is unsure about* CMT*. All four columns and key* VT *form input for our tool, and the team selects Top-down, Breadth-first as traversal strategy. Fig. 2 shows the column lattice and sequence of potential keys traversed. Minimal keys are in dark blue and form the target set R, keys in light blue are supersets of minimal keys, keys in orange are meaningless, and keys in red are maximal among the meaningless. After our experts dismiss VMC as meaningless, none of its subsets needs traversal. Neither* TVM*,* TVC*,* $TVMC$ *nor* T *need questioning as* VT *was classified as minimal key in the input. Fig. 3 shows the tool's interface before the last question is answered. Answering "No" will add* CT *as another meaningful minimal key. Question 4 asks if it is possible that this key could be violated (eg by the Interview Sample), hence answering "No" means the key is meaningful. The tool also shows a relation that is Armstrong for the current set of minimal keys (*MT *and* TV*), that is, it satisfies a key if and only if it is a superset for some of these minimal keys (hence,* CT *is violated). The final result is* $R = \{VT, CT, MT\}$*. Compared to* $S = \{VT, CMT\}$*, use of the Dragon's Dojo has correctly removed the false positive* CMT *and added the false negatives* CT *and* MT*.*



(a) Simplified Methodology **(b) Interface for Input**

Figure 4: Simplified Methodology and Interface for Input

Sec. 2 explains the architecture, methodology and features of the Dragon's Dojo, positions its novelty and highlights the impact. Sec. 3 details what the audience will see: an overview of the tool, scenarios showcasing traversal strategies, predetermined columns and keys, functionality that extracts the expertise of experts, and a competition engaging the audience.

2 THE DRAGON'S DOJO

We will describe the architecture and interview process of the Dragon's Dojo, followed by highlighting its novelty and impact.

2.1 Methodology and Architecture

We start with a simplified methodology with two steps: the input and interview process. For the input, experts select one of four traversal strategies: Top-down, Breadth-first (TB), Top-down, Depth-first (TD), Bottom-up, Breadth-first (BB), or Bottom-up, Depth-first (BD). Subsequently, they specify the column names (schema attributes) for which minimal keys need to be identified. They may provide domain values for each column and any predetermined minimal keys. The interview process computes yes/no questions the team answers collectively until the set R of meaningful minimal keys is determined. The sequencing of questions depends on the input and previous answers. The number of questions generated is minimal for the input. Fig. 4a illustrates the simplified methodology, while Fig. 4b and Fig. 3 show the interface for i) acquiring input from the team, and ii) answering questions by the team, respectively.

For example, [Question 4] in Fig. 3 validates the potential key $\{Chair, Time\}$ by asking the team if it is possible that there are two records with values matching on *Chair* and *Time*, like in the Interview Sample provided. Answering "No" means the key is meaningful and, as indicated in Fig. 2, the input and previous answers establish $\{Chair, Time\}$ as remaining minimal key. Since TV is a predetermined minimal key, T is meaningless and TVM, TVC and $TVMC$ are superkeys of TV, while a "Yes" answer to the validation of VMC has dismissed its meaningfulness and that of all its subsets.

As illustrated in Fig. 3 already, the Dragon's Dojo supports a much richer methodology that incorporates previous work on Armstrong databases to access the expertise of domain experts as well. This methodology is illustrated in Fig. 5, with the simplified part of the overall methodology highlighted in red. Any set C of minimal keys can be represented in form of an Armstrong database over the input schema, satisfying a key if and only if it is implied by C.

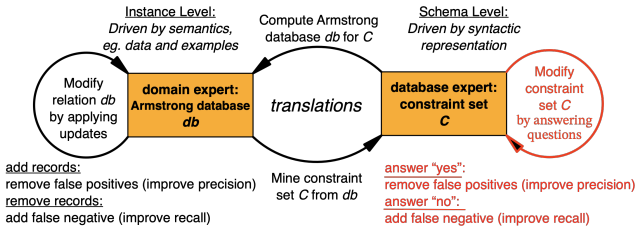
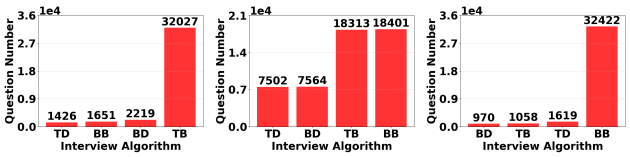


Figure 5: Methodology to Acquire the Set of Meaningful Database Keys with Perfect Precision and Recall



(a) Bottom **(b) Middle** **(c) Top**

Figure 6: Number of questions for randomly distributed minimal keys on a schema with $n = 15$ columns, based on which layer their average size k belongs: Bottom ($k \in [0, n/3]$), Middle ($k \in (n/3, 2n/3]$) and Top ($k \in (2n/3, n]$).

Led by the domain experts, the team may modify the Armstrong database by removing/adding records or updating values. Using the discovery of keys from data, updates to Armstrong relations can be translated to changes in the set C of minimal keys. Experts benefit from two different points of view, each emphasizing their strength in expertise: database experts know the schema level well, while domain experts understand the instance level. Translations map one view into the other. Experts can overturn their decisions, highlighting the challenge of decision-making. Fig. 3 shows an Armstrong relation for the minimal keys MT and TV.

2.2 Novelty and Impact

The Dragon's Dojo is the first tool that computes interview questions to help human experts classify keys as meaningful. It complements existing methodologies that discover keys from data [1], implemented in tools like Metanome [2, 8, 9] and DataViadotto [4, 5], and computing Armstrong relations [6, 7], implemented in tools like DataProf [10]. Data profiling only returns recommendations for minimal keys without classifying them as meaningful, which only human experts can do. Tools that compute Armstrong relations do not systematically traverse the search space to guide human experts in identifying meaningful keys.

Regarding impact, getting to know how entities can be identified uniquely is central to all data processing, management and analytical tasks. Direct benefits include: (1) Effective data organization because minimal keys (a) provide pathways to critical data elements within the database, (b) constitute well-defined reference points that facilitate the joining of different data tables, (c) keep data utilization effective in the presence of missing and inconsistent data; (2) Enhanced data quality since minimal keys (a) enforce entity integrity, ensuring that entities within the database are unique and accurately represented, (b) can identifying duplicate IDs associated with the same entity, thereby eliminating redundancy and enhancing data accuracy, (c) help implement techniques for imputing missing data; (3) Better performance since minimal keys (a) leverage unique indexes to facilitate faster data access, improving the overall performance of queries and data retrieval processes, (b) can optimize join types and other queries to ensure that data retrieval is as efficient as possible, (c) include business keys that enhance trust in the data, providing stakeholders with confidence in the integrity and applicability of the information.

Utilizing the Dragon's Dojo, organizations cannot only arrange their data more effectively but also significantly enhance its quality and performance, leading to better decision-making and outcomes.

3 DEMONSTRATION

We outline five scenarios the audience will experience.

3.1 Scenario 1: Real-world Minimal Keys

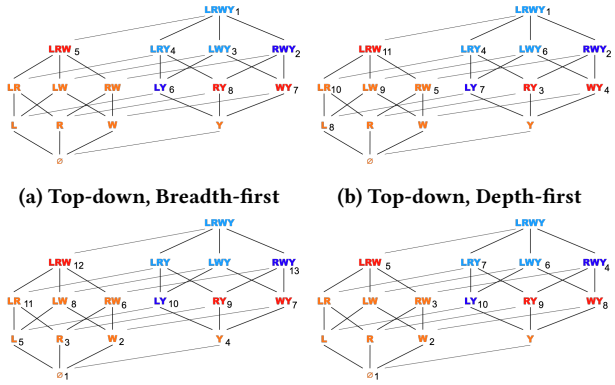
For the first scenario we will use a poster to explain the problem, illustrate its challenges and scale, show which traversal strategies work well for which instances, and discuss techniques to handle real-world distributions of minimal keys. Detailed background material can be found in a technical paper on our artifact page[1].

For example, Fig. 6 shows how the distribution of minimal keys affects the number of questions generated by the traversal strategies. For bottom (top) layers, TB (BB, respectively) is prohibitively expensive. For middle layers, breadth-first traversals are generally costly, while depth-first strategies perform significantly better. In general, the largest families of minimal keys over a schema with n columns have size $\binom{n}{\lfloor n/2 \rfloor}$, so even the solution size is worst-case exponential in the number of columns.

The challenge is illustrated on the SeriesPost table of the public hockey dataset[2], which models playoff series between teams in leagues. The table has 13 columns and no keys exist on the schema. Following in-depth analysis, the ground truth for the minimal keys is {*year*, *series*}, {*year*, *round*, *winner*} and {*year*, *loser*}. With all 13 columns as input, our traversal strategies require the following numbers of questions to obtain the ground truth: BB-4867, BD-3327, TD-3279, and TB-3331. These numbers quantify the challenge: human experts must consider these many keys to solve the problem.

In practice, we reduce the input to columns feasible to occur in some key (eg statistics like *goals* is unreasonable). The expert team does this manually or by using some mining tool, if data is available. We will show four viable keys returned by the DataViadotto tool [4]: {*year*, *series*}, {*year*, *round*, *winner*}, {*year*, *round*, *loser*} and {*year*, *winner*, *loser*}. Using columns *year*, *series*, *round*, *winner* and *loser* as input, the Dragon's Dojo obtains the ground truth by BB-22, BD-16, TD-16, and TB-16, which is very feasible. Mining has only a precision of 1/2, and recall of 1/3, so using the Dragon's Dojo improves precision by 1/2 and recall by 2/3. Approximate mining of keys would also return {*year*, *loser*} with less than 1% dirty data, but the point is that we need human experts to classify keys as meaningful, irrespective whether they are exact or approximate.

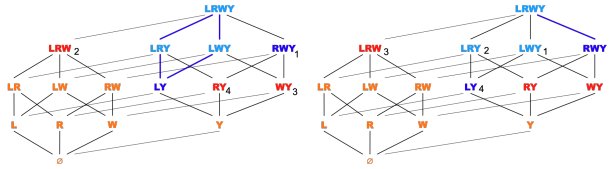
[1] https://github.com/zhuoxingzhang/DD
[2] https://relational.fel.cvut.cz/dataset/Hockey



(a) Top-down, Breadth-first (b) Top-down, Depth-first

(c) Bottom-up, Breadth-first (d) Bottom-up, Depth-first

Figure 7: Traversal strategies on `SeriesPost`



(a) Using LY (b) Using RWY

Figure 8: Fewer Interview Questions by Predetermined Keys

3.2 Scenario 2: Interview Traversals

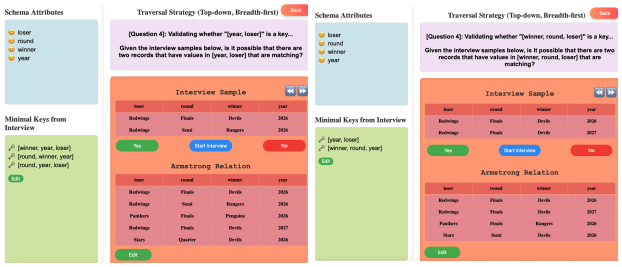
Our second scenario showcases the impact of different traversal strategies on the number of questions that the human experts are required to answer. Fig. 7 illustrates the sequences of questions generated by the different traversal strategies. For simplicity, we use columns *L(oser)*, *R(ound)*, *W(inner)*, and *Y(ear)*. Here, the audience will develop a deeper understanding how different traversal strategies can lead to faster outcomes in practical scenarios. For the specific distribution of minimal keys in this example, TB works well as it detects both minimal keys and maximal anti-keys early.

3.3 Scenario 3: Using Predetermined Keys

Another scenario highlights how the Dragon's Dojo uses already confirmed results to lower the complexity of interviews. For our example, we add either LY or RWY as predetermined minimal key to the input. Fig. 8 illustrates how the additional input lowers the number of questions in each case.

3.4 Scenario 4: Modifying Sample Data

As shown in Fig. 5, human experts can use interview questions and Armstrong relations. Fig. 9 illustrates how the team decides to remove an entire row from the current Armstrong relation, rather than answering the next interview question. In particular, Fig. 9a shows how the Dragon's Dojo questions the potential key {*year*, *loser*}, given the current set of three minimal keys. Fig. 9b shows the impact of removing the second row in the Armstrong relation for the three minimal keys: the modified Armstrong relation now satisfies the two minimal keys {*year*, *loser*} and {*winner*, *round*, *year*}.



(a) Before Removing Row Two (b) After Row Removal

Figure 9: Before and After Updating the Armstrong Relation

In effect, the two previously minimal keys have been replaced by the new minimal key {*year*, *loser*}, which constitutes the major difference between the mining results and the ground truth.

3.5 Scenario 5: Competing in the Dragon's Dojo

Having gained a thorough understanding through the previous scenarios, the audience will be invited to enter the Dragon's Dojo for a competition. Using a public web interface, they need to find the correct set of minimal keys for our Calendar schema as input. The audience can take a photo of their score and send it to us. Teams closest to the correct answer win, using precision and recall. In case of ties, all teams with the fewest questions required win.

4 CONCLUSION

Understanding which database keys govern an application domain is a necessity for effective data management. Our demo showcases the first computational tool for helping human experts classify database keys as meaningful. It quantifies the challenge of having to consider an overwhelming number of candidates, and shows how the tool provides an effective methodology for utilizing interviews and samples to address the problem.

REFERENCES

[1] Ziawasch Abedjan, Lukasz Golab, Felix Naumann, and Thorsten Papenbrock. 2018. *Data Profiling*. Morgan & Claypool Publishers.

[2] Johann Birnick, Thomas Bläsius, Tobias Friedrich, Felix Naumann, Thorsten Papenbrock, and Martin Schirneck. 2020. Hitting Set Enumeration with Partial Information for Unique Column Combination Discovery. *Proc. VLDB Endow.* 13, 11 (2020), 2270–2283.

[3] David W. Embley. 2018. Key. In *Encyclopedia of Database Systems, Second Edition*.

[4] Henning Koehler and Sebastian Link. 2025. Mining Meaningful Keys and Foreign Keys with High Precision and Recall. *Proc. VLDB Endow.* 18, 12 (2025), 5363–5366.

[5] Henning Koehler and Sebastian Link. 2025. Orthogonal Keys: High Precision and Recall for Mining Database Keys From Inconsistent and Incomplete Relations. *IEEE Trans. Knowl. Data Eng.* 37, 11 (2025), 6550–6561.

[6] Heikki Mannila and Kari-Jouko Räihä. 1986. Design by Example: An Application of Armstrong Relations. *J. Comput. Syst. Sci.* 33, 2 (1986), 126–141.

[7] Fabien De Marchi and Jean-Marc Petit. 2007. Semantic sampling of existing databases through informative Armstrong databases. *Inf. Syst.* 32, 3 (2007), 446–457.

[8] Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. 2015. Data Profiling with Metanome. *Proc. VLDB Endow.* 8, 12 (2015), 1860–1863.

[9] Eduardo H. M. Pena, Eduardo C. de Almeida, and Felix Naumann. 2019. Discovery of Approximate (and Exact) Denial Constraints. *Proc. VLDB Endow.* 13, 3 (2019), 266–278.

[10] Ziheng Wei and Sebastian Link. 2018. DataProf: Semantic Profiling for Iterative Data Cleansing and Business Rule Acquisition. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD*. 1793–1796.